Statistics 582, Problem Set 3 Solutions

Wellner; 1/30/98

1. Suppose that $X, X_1, \ldots, X_n$ are i.i.d. Weibull(α_0, β_0) (if X has the Weibull(θ) distribution where $\theta = (\alpha, \beta)$, then $1 - F_\theta(x) = P_\theta(X > x) = \exp(-(x/\alpha)^\beta)$ for $x \geq 0$). Recall that the MLE $\widehat{\alpha}$ of α is given by

$$\widehat{\alpha} = \left\{ \frac{1}{n} \sum_{i=1}^n X_i^{\widehat{\beta}} \right\}^{1/\widehat{\beta}}$$

where $\widehat{\beta}$ is the MLE of β. As a simpler alternative to maximum likelihood, I propose to use the alternative estimator $\overline{\beta}$ of β obtained from the slope of an ordinary least squares fit of a Weibull Q-Q plot, and then estimate α by

$$\overline{\alpha} = \left\{ \frac{1}{n} \sum_{i=1}^n X_i^{\overline{\beta}} \right\}^{1/\overline{\beta}} .$$

A. Suppose that $\overline{\beta}_n \to_p \beta_0$ is known. Show that $\overline{\alpha}_n \to_p \alpha_0$. [Hint: use a uniform strong law of large numbers.]
B. Show that $\overline{\alpha}_n$ is a "pseudo-MLE" in the sense that $\overline{\alpha}_n$ maximizes $l_n(\alpha, \overline{\beta}_n)$.

Solution: Fix $\delta > 0$ (small). The family of functions $\mathcal{F} = \{f(x, \beta) = x^\beta : \beta \in [\beta_0 - \delta, \beta_0 + \delta]\}$ are indexed by the compact set $[\beta_0 - \delta, \beta_0 + \delta]$, are continuous in β for every $x \geq 0$, and are bounded by

$$\sup_{\beta \in [\beta_0 - \delta, \beta_0 + \delta]} |f(x, \beta)| = x^{\beta_0 + \delta} \vee x^{\beta_0 - \delta} \leq x^{\beta_0 + \delta} + x^{\beta_0 - \delta} \equiv F(x)$$

which satisfies $E_0 F(X) < \infty$ if $\delta < 2\beta_0$. Thus by theorem 4.4.1 (of the section 4 revision) the uniform strong law of large numbers holds for $\mathcal{F}$:

$$\sup_{\beta : |\beta - \beta_0| \leq \delta} |\mathbb{P}_n f(\cdot, \beta) - P_0 f(\cdot, \beta)| \to_{a.s.} 0 .$$

Since $\overline{\beta}_n \to_{a.s.} \beta_0$, $\overline{\beta}_n \in [\beta_0 - \delta, \beta_0 + \delta]$, with probability 1 for n sufficiently large, and it follows from the uniform strong law law of large numbers (Theorem 1, section 4.4 revision) together with continuity of $\mu(\beta) \equiv E_0 f(X, \beta)$ that

$$\begin{aligned} \overline{\alpha}_n^{\overline{\beta}_n} &= \frac{1}{n} \sum_{i=1}^n X_i^{\overline{\beta}_n} \\ &\to_{a.s.} E_0 f(X, \beta_0) = \alpha_0^{\beta_0} . \end{aligned}$$

But now

$$\overline{\alpha}_n = \{\overline{\alpha}_n^{\overline{\beta}_n}\}^{1/\overline{\beta_n}} = g(\overline{\alpha}^{\overline{\beta}_n}, \overline{\beta_n})$$

where $g(u,v) \equiv u^{1/v}$ is continuous and $(\overline{\alpha}^{\overline{\beta}_n}, \overline{\beta_n}) \to_{a.s.} (\alpha_0^{\beta_0}, \beta_0)$. Hence by the continuous mapping theorem

$$\overline{\alpha}_n = g(\overline{\alpha}_n^{\overline{\beta}_n}, \overline{\beta_n}) \to_{a.s.} g(\alpha_0^{\beta_0}, \beta_0) = \alpha_0 \,.$$

B. The log-likelihood is

$$l_n(\alpha, \beta) = n \log(\beta/\alpha) + (\beta - 1) \sum_{i=1}^{n} \log(X_i/\alpha) - \sum_{i=1}^{n} \left(\frac{X_i}{\alpha}\right)^{\beta},$$

and hence

$$\begin{aligned} l_n(\alpha, \overline{\beta}_n) &= n \log(\overline{\beta}_n/\alpha) + (\overline{\beta}_n - 1) \sum_{i=1}^{n} \log(X_i/\overline{\alpha}) - \sum_{i=1}^{n} \left(\frac{X_i}{\alpha}\right)^{\overline{\beta}_n} \\ &= -n\overline{\beta}_n \log \alpha - \frac{\sum X_i^{\overline{\beta}_n}}{\alpha^{\overline{\beta}_n}} + \text{constant in } \alpha \\ &= -n \log \eta - \frac{\sum X_i^{\overline{\beta}_n}}{\eta} + \text{constant in } \alpha \text{ and } \eta \end{aligned}$$

where $\eta \equiv \alpha^{\overline{\beta}_n}$. This is easily seen to be maximized by

$$\overline{\eta} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i^{\overline{\beta}_n}$$

and hence

$$\overline{\alpha}_n = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^{\overline{\beta}_n} \right\}^{1/\overline{\beta}_n}$$

as claimed. Thus $\overline{\alpha}_n$ is a pseudo-MLE of α.

2. Human beings can be classified into one of four blood groups (phenotypes) O, A, B, AB. The inheritance of blood groups is controlled by three genes, O, A, B, of which O is recessive to A and B. If r, p, q are the gene probabilities in the population of O, A, B respectively, then the probabilities of the six possible combinations (genotypes) in random mating (where two individuals drawn at random from the population contribute one gene each) are shown in the following table:

Phenotype	Genotype	probability
O	OO	r^2
A	AA	p^2
A	AO	$2rp$
B	BB	q^2
B	BO	$2rq$
AB	AB	$2pq$

We observe among N individuals the phenotype frequencies N_O, N_A, N_B, N_{AB}, and wish to estimate the gene probabilities from such data. A simple approach is to regard the observations as incomplete, the complete data set being the genotype frequencies N_{OO}, N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}.
A. Derive the EM algorithm for estimation of (p, q, r).
B. Estimate (p, q, r) from $N_O = 176$, $N_A = 182$, $N_B = 60$, $N_{AB} = 17$.
C. Estimate the covariance matrix of the estimator $(\hat{p}, \hat{q}, \hat{r})$.

Solution: A. The complete data is $\underline{N} \equiv (N_{OO}, N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB})$ with multinomial distribution $\text{Mult}_6(N; (r^2, p^2, 2rp, q^2, 2rq, 2pq))$. Thus

$$P(\underline{N} = \underline{n}) = \frac{N!}{n_{OO}!n_{AA}!n_{AO}!n_{BB}!n_{BO}!n_{AB}!} \cdot p^{2n_{AA}+n_{AO}+n_{AB}}\, q^{2n_{BB}+n_{BO}+n_{AB}}\, r^{2n_{OO}+n_{AO}+n_{BO}}\, 2^{n_{AO}+n_{BO}+n_{AB}}.$$

This is proportional to a $\text{Mult}_3(2N; (p, q, r))$ distribution, and hence the MLE's based on the complete data are

$$(\hat{p}, \hat{q}, \hat{r}) = \frac{1}{2N}(2N_{AA} + N_{AO} + N_{AB}, 2N_{BB} + N_{BO} + N_{AB}, 2N_{OO} + N_{AO} + N_{BO}).$$

This forms the basis of the "M - step"of an E-M algorithm. The incomplete data Y is $(N_A, N_B, N_O, N_{AB}) = (N_{AA} + N_{AO}, N_{BB} + N_{BO}, N_{OO}, N_{AB})$; thus

$$(N_{AA}|Y) = (N_{AA}|N_A) \sim \text{Binomial}(N_A, \frac{p^2}{p^2 + 2rp}), \qquad E(N_{AA}|Y) = N_A \frac{p}{p + 2r},$$

$$(N_{AO}|Y) = (N_{AO}|N_A) \sim Binomial(N_A, \frac{2rp}{p^2 + 2rp}), \qquad E(N_{AO}|Y) = N_A \frac{2r}{p + 2r},$$

$$(N_{BB}|Y) = (N_{BB}|N_B) \sim Binomial(N_B, \frac{q^2}{q^2 + 2rq}), \qquad E(N_{BB}|Y) = N_B \frac{q}{q + 2r},$$

$$(N_{BO}|Y) = (N_{BO}|N_B) \sim Binomial(N_B, \frac{2rq}{q^2 + 2rq}), \qquad E(N_{BO}|Y) = N_B \frac{2r}{q + 2r}.$$

This gives the basis of the "E - step" for an E - M algorithm. Hence, starting from $(\hat{p}^{(0)}, \hat{q}^{(0)}, \hat{r}^{(0)}) = (1/3, 1/3, 1/3)$ say, we take

$$(\hat{p}^{(m+1)}, \hat{q}^{(m+1)}) = \frac{1}{2N}(2\hat{N}_{AA}^{(m)} + \hat{N}_{AO}^{(m)} + N_{AB}, 2\hat{N}_{BB}^{(m)} + \hat{N}_{BO}^{(m)} + N_{AB}),$$

$$\hat{r}^{(m+1)} = 1 - \hat{p}^{(m+1)} - \hat{q}^{(m+1)}$$

where

$$\hat{N}_{AA}^{(m)} \equiv N_A \frac{\hat{p}^{(m)}}{\hat{p}^{(m)} + 2\hat{r}^{(m)}}, \qquad \hat{N}_{AO}^{(m)} \equiv N_A - \hat{N}_{AA}^{(m)},$$

$$\hat{N}_{BB}^{(m)} \equiv N_B \frac{\hat{q}^{(m)}}{\hat{q}^{(m)} + 2\hat{r}^{(m)}}, \qquad \hat{N}_{BO}^{(m)} \equiv N_B - \hat{N}_{BB}^{(m)}.$$

B. For the given data, the E - M algorithm in A yields:

Iteration	$\hat{p}^{(m)}$	$\hat{q}^{(m)}$
0	.333	.333
1	.298	.111
2	.271	.094
3	.266	.093
4	.265	.093
5	.264	.093
6	.264	.093

Thus the estimator is $(\hat{p}, \hat{q}, \hat{r}) = (.264, .093, .642)$.

C. The likelihood of the observations $(N_A, N_B, N_O, N_{AB}) = (N_{AA} + N_{AO}, N_{BB} + N_{BO}, N_{OO}, N_{AB})$ is

$$\begin{aligned} l_N(p,q) &= N_A \log(p^2 + 2p(1-p-q)) \\ &\quad + N_B \log(q^2 + 2q(1-p-q)) \\ &\quad + N_O \log(1-p-q)^2 + N_{AB} \log(2pq). \end{aligned}$$

Thus

$$\begin{aligned} -\frac{\partial^2}{\partial p^2} l_N(p,q) &= 2N_A\{\frac{1}{2p-p^2-2pq} + \frac{2(1-p-q)^2}{(2p-p^2-2pq)^2}\} \\ &\quad + N_B \frac{4q^2}{(2q-q^2-2pq)^2} \\ &\quad + \frac{N_{AB}}{p^2} + \frac{2N_O}{(1-p-q)^2}, \end{aligned}$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial p \partial q} l_N(p,q) &= 2N_A\{\frac{1}{2p-p^2-2pq} - \frac{2p(1-p-q)}{(2p-p^2-2pq)^2}\} \\
&\quad + 2N_B\{\frac{1}{2q-q^2-2pq} - \frac{4q^2}{(2q-q^2-2pq)^2}\} \\
&\quad + \frac{2N_O}{(1-p-q)^2}\,,
\end{aligned}
$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial q^2} l_N(p,q) &= N_A \frac{4p^2}{(2p-p^2-2pq)^2} \\
&\quad + 2N_B\{\frac{1}{2q-q^2-2pq} + \frac{2(1-p-q)^2}{(2q-q^2-2pq)^2}\} \\
&\quad + \frac{N_{AB}}{q^2} + \frac{2N_O}{(1-p-q)^2}\,.
\end{aligned}
$$

Since

$$
\begin{aligned}
E(N_A) &= N(p^2 + 2p(1-p-q))\,, \\
E(N_B) &= N(2q-q^2-2pq)\,, \\
E(N_{AB}) &= N(2pq),
\end{aligned}
$$

and

$$
E(N_O) = N(1-p-q)^2\,,
$$

it follows that

$$
I_{11}(p,q) = 2N\{1 + \frac{2r^2}{2p-p^2-2pq} - \frac{2q^2}{2q-q^2-2pq} + \frac{q}{p} + 1\},
$$

$$
I_{12}(p,q) = 2N\{2 - \frac{2p(1-p-q)}{(2p-p^2-2pq)^2} - \frac{2q(1-p-q)}{(2q-q^2-2pq)^2} + 1\}\,,
$$

$$
I_{22}(p,q) = 2N\{1 + \frac{2r^2}{2q-q^2-2pq} - \frac{2p^2}{2p-p^2-2pq} + \frac{p}{q} + 1\}
$$

and hence the estimated Fisher information matrix is

$$
\hat{I}(p,q) = \begin{pmatrix} 5.063 & 1.793 \\ 1.793 & 12.182 \end{pmatrix}
$$

so that

$$
\hat{I}^{-1}(p,q) = \frac{1}{2N}\begin{pmatrix} .208 & -.003 \\ -.003 & .087 \end{pmatrix}\,.
$$

Furthermore, since $\hat{r} = 1 - \hat{p} - \hat{q}$,

$$
Var(\hat{r}) = Var(\hat{p}) + Var(\hat{q}) + 2Cov(\hat{p},\hat{q})\,,
$$

$$Cov(\hat{p},\hat{r}) = -Var(\hat{p}) - Cov(\hat{p},\hat{q})\,,$$

$$Cov(\hat{q},\hat{r}) = -Var(\hat{q}) - Cov(\hat{p},\hat{q})\,;$$

and hence we estimate $Cov(\hat{p},\hat{q},\hat{r})$ by

$$\widehat{Cov}(\hat{p},\hat{q},\hat{r}) = \begin{pmatrix} .000240 & -.000035 & -.000205 \\ -.000035 & .000095 & -.000060 \\ -.000205 & -.000060 & .000265 \end{pmatrix}.$$

3. Suppose that the "complete data" X is given by three independent multinomial random vectors,

$$N(1) \equiv (N_{ij}(1) : i = 1,\ldots,r; j = 1,\ldots,s) \sim \text{Mult}_{rs}(n_1; p = (p_{ij}, i = 1,\ldots,r, j = 1,\ldots,s),$$

$$N(2) \equiv (N_{ij}(2) : i = 1,\ldots,r; j = 1,\ldots,s) \sim \text{Mult}_{rs}(n_2; p = (p_{ij}, i = 1,\ldots,r, j = 1,\ldots,s),$$

$$N(3) \equiv (N_{ij}(3) : i = 1,\ldots,r; j = 1,\ldots,s) \sim \text{Mult}_{rs}(n_3; p = (p_{ij}, i = 1,\ldots,r, j = 1,\ldots,s)\,.$$

Suppose that the "incomplete data" Y consists of $N(1), (N_{i\cdot}(2) : 1 \le i \le r), (N_{\cdot j}(3) : 1 \le j \le s)$.
A. What are the distributions of $N(1), (N_{i\cdot}(2) : 1 \le i \le r)$ and $(N_{\cdot j}(3) : 1 \le j \le s)$?
B. Find the conditional distribution(s) of X given Y.
C. Suggest an EM - algorithm for estimation of p.

Solution: A. By elementary considerations,

$$(N_{i\cdot}(2) : 1 \le i \le r) \sim Mult_r(n_2; (p_{i\cdot} : 1 \le i \le r))$$

and

$$(N_{\cdot j}(3) : 1 \le j \le s) \sim Mult_s(n_3; (p_{\cdot j} : 1 \le j \le s)).$$

B. First note that if

$$(N_{ij}) \sim \text{Mult}_{rs}(n; (p_{ij})),$$

then

$$(N_{i\cdot}) \sim \text{Mult}_r(n; (p_{i\cdot}))$$

as in A (since the components of $(N_{i\cdot})$ give the number of times outcome i occurred in n independent trials with probability $p_{i\cdot}$ on each trial). Furthermore

$$((N_{ij})|(N_{i\cdot})) \sim \prod_{i=1}^{r} Mult_s(N_{i\cdot}; (p_{ij}/p_{i\cdot}))\,. \tag{0.1}$$

(0.1) can be proved most easily by direct calculation of the conditional distribution:

$$\begin{aligned} & P(N_{ij} = k_{ij}, i = 1, \ldots, r, j = 1, \ldots, s \Big| N_{i\cdot} = k_{i\cdot}, i = 1, \ldots, r) \\ = \; & n! \prod_{i=1}^{r} \prod_{j=1}^{s} \frac{p_{ij}^{k_{ij}}}{k_{ij}!} / n! \prod_{i=1}^{r} \frac{p_{i\cdot}^{k_{i\cdot}}}{k_{i\cdot}!} \\ = \; & \prod_{i=1}^{r} \left\{ k_{i\cdot}! \prod_{j=1}^{s} \frac{(p_{ij}/p_{i\cdot})^{k_{ij}}}{k_{ij}!} \right\} \end{aligned}$$

on the set $k_{i\cdot} = \sum_{j=1}^{s} k_{ij}$, $i = 1, \ldots, r$. The terms inside the first product are just the $\mathrm{Mult}_s(k_{i\cdot}; (p_{ij}/p_{i\cdot}))$ probabilities.

Hence conditional on $(N_{i\cdot}(2) : 1 \le i \le r)$ the vectors $(N_{ij}(2) : 1 \le j \le s)$, $i = 1, \ldots, r$ are independent with $(N_{ij}(2) : 1 \le j \le s)|N_{i\cdot} \sim \mathrm{Mult}_s(N_{i\cdot}; (p_{ij}/p_{i\cdot}; j = 1, \ldots, s))$. Similarly, conditional on $(N_{\cdot j}(3) : 1 \le j \le s)$ the vectors $(N_{ij}(3) : 1 \le i \le r)$, $j = 1, \ldots, s$ are independent with $(N_{ij}(3) : 1 \le i \le r)|N_{\cdot j} \sim \mathrm{Mult}_r(N_{\cdot j}; (p_{ij}/p_{\cdot j}; i = 1, \ldots, r))$.

C. If we had the complete data $N_{ij}(1), N_{ij}(2), N_{ij}(3)$ for all i, j, then $N_{ij} \equiv N_{ij}(1) + N_{ij}(2) + N_{ij}(3)$ has a multinomial distribution with number of trials $n \equiv n_1 + n_2 + n_3$, and hence the MLE $\underline{\hat{p}} = (\hat{p}_{ij})$ of $\underline{p} = (p_{ij})$ is given by

$$\hat{p}_{ij} = \frac{N_{ij}}{n} = \frac{N_{ij}(1) + N_{ij}(2) + N_{ij}(3)}{n_1 + n_2 + n_3}.$$

This is the basis of the "M - step" of an E-M algorithm. But from B it follows that

$$E(N_{ij}(2)|N_{i\cdot}(2)) = N_{i\cdot}(2)\frac{p_{ij}}{p_{i\cdot}}, \qquad E(N_{ij}(3)|N_{\cdot j}(3)) = N_{\cdot j}(3)\frac{p_{ij}}{p_{\cdot j}}.$$

This is the basis of the "E - step" of an E-M algorithm. Thus, for some reasonable preliminary estimator like $\underline{\hat{p}}^{(0)} \equiv (\hat{p}_{ij}^{(0)}) = (N_{ij}(1)/n)$, a natural E - M algorithm is defined by

$$\hat{p}_{ij}^{(m+1)} = \frac{N_{ij}(1) + \widehat{N}_{ij}^{(m)}(2) + \widehat{N}_{ij}^{(m)}(3)}{n_1 + n_2 + n_3}$$

where

$$\widehat{N}_{ij}^{(m)}(2) \equiv N_{i\cdot}(2)\frac{\hat{p}_{ij}^{(m)}}{\hat{p}_{i\cdot}^{(m)}}, \qquad \widehat{N}_{ij}^{(m)}(3) \equiv N_{\cdot j}(3)\frac{\hat{p}_{ij}^{(m)}}{\hat{p}_{\cdot j}^{(m)}}.$$